Ahrii Kim[1]

[1]Independent Researcher

August 26, 2025

Context is Ubiquitous, but Rarely Changes Judgments: Revisiting Document-Level MT Evaluation

Ahrii Kim
Independent Researcher / Seoul, Korea
ahriikim@gmail.com

Abstract

As sentence-level performance in modern Machine Translation (MT) models reaches a plateau where differences are minimal, there is a growing need for robust document-level evaluation methods. We present a reproducible human evaluation protocol that is structured upon the FALCON framework (Kim, 2025) encompassing pragmatic features. With professional translators as annotators, we investigate the sources of low inter-annotator agreement and identify the primary contributing factors.

To address these challenges and align with human values, we propose a comprehensive annotation-rating methodology referred to as **H-FALCON**. Our experiment shows that, while perfect annotator consensus remains elusive, the proposed scoring scheme achieves equal or higher correlations with traditional sentence-level metrics. Linear regression analysis further reveals that contextual information is inherent in all sentences—contrary to the belief that only a subset requires it—and that previous estimates such as "n % of sentences require context" stem from flawed calculations. Context contributes approximately 10% to the variance of the holistic score in our evaluation, highlighting its universal yet limited influence on the MT evaluation. Codes will be released.

1 Introduction

The conventional approach to automatic machine translation evaluation has predominantly concentrated on sentence-level analysis, emphasizing lexical overlap or n-gram similarity, as exemplified by metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and chrF (Popović, 2015). Additionally, it has considered semantic similarities through embedding-based metrics, including BERTScore (Zhang et al., 2019) and COMET (Rei et al., 2020). In recent developments, LLM-based (large language model) metrics, such as XCOMET (Guerreiro et al., 2024) and MetaMetrics (Anugraha et al., 2024), have demonstrated enhanced alignment with human values. Nevertheless, a significant limitation of these methodologies resides in their restriction to sentence-level analysis, thereby failing to capture discourse phenomena such as cohesion, coreference, consistency, and pragmatic adequacy. Several automatic metrics have been proposed to address this issue (Jwalapuram et al. 2021; Zhao et al. 2023; Jiang et al. 2022); however, these metrics face challenges related to their reliability and reproducibility.

The execution of human evaluation at the document level also encounters considerable challenges, primarily due to the heightened cognitive efforts placed upon evaluators (Läubli et al. 2018; Bawden et al. 2018), subsequently leading to prolonged annotation duration and decreased inter-annotator agreement (IAA) (Graham et al., 2017). Furthermore, significant variability exists among evaluation protocols with respect to context length, annotation granularity, or provided guidelines (Hardmeier et al. 2015; Kocmi et al. 2022). Quantifying context-dependent phenomena is also challenging, and relying only on counting overt discourse markers can lead to errors, underestimating the importance of context in translation (Voita et al. 2019; Castilho 2022).

The confluence of these factors results in document-level evaluation being both methodologically intricate and resource-demanding, thereby impeding its extensive integration into MT research and practice. Nevertheless, the recent framework FALCON (Kim, 2025) seeks to bridge this gap by incorporating pragmatic features into a structured document-level evaluation. Regrettably, their human evaluation aspect is excessively minimized in scope, and its components remain unverified regarding reproducibility and reliability. We therefore put forth H-FALCON, a reproducible protocol for document-level human evaluation, validated with professional translators. Our contributions are:

- **First systematic reliability study** of FALCON, identifying sources of inter-annotator variation.
- **Comprehensive meta-evaluation** of FALCON across diverse proprietary models, revealing its limitations.
- **Introduction of H-FALCON**, unifying annotation and rating with pragmatic features.
- **Evidence** that contextual information is inherent in all sentences, challenging prior claims that only a subset requires it.
- **Quantitative insight** showing context explains a bounded yet consistent share (about 10%) of variance in holistic MT scores.

2 Related Works

Document-level evaluation is not simply a scaled-up version of sentence-level evaluation; it captures translation phenomena from extended context. Phenomena such as coreference resolution, lexical cohesion, discourse connective use, and pragmatic intent often require evaluators to consider multiple sentences or entire documents rather than isolated segments (Thai et al. 2022; Dahan et al. 2024). This broader perspective can reveal quality differences between MT systems that appear statistically indistinguishable at the sentence level (Sharma and Sridhar, 2025). This section will elucidate the efforts made in both manual (§ 2.1) and automatic evaluation (§ 2.2, § 2.3) research to address document-level phenomena.

2.1 For Manual Evaluation

Frameworks such as MQM (Multidimensional Quality Metrics; Lommel et al. 2014) and TAUS DQF (Dynamic Quality Framework; Valli 2015) provide detailed error taxonomies but are typically adapted in ad-hoc ways for document-level use, leading to inconsistent application and reduced IAA (Graham et al., 2017). Shared tasks such as DiscoMT (Workshop on Discourse in Machine Translation; Hardmeier et al. 2015) and WMT campaigns (Conference on Machine Translation; Kocmi et al. 2022; Kocmi et al. 2023; Kocmi et al. 2024) have introduced protocols where annotators evaluate entire documents or extended context windows to capture cross-sentence dependencies. However, reproducibility and reliability remain persistent challenges, especially when pragmatic adequacy is involved.

2.2 For Automatic Evaluation

On the machine side, several automatic metrics have been adapted to better capture discourse and context in MT evaluation. DiscoScore (Zhao et al., 2023) explicitly models discourse relations and coreference chains to assess cohesion and coherence. BlonDE (Jiang et al., 2022) incorporates lexical, syntactic, semantic, and discourse-level features to evaluate narrative and dialogic text, making it suitable for context-rich MT outputs. Doc-COMET (Vernikos et al., 2022) extends the COMET framework (Rei et al., 2020) to accept document-level inputs, leveraging contextual embeddings to evaluate translations in their broader discourse environment. While these approaches represent important progress toward automated document-level evaluation, they typically focus on one or two discourse aspects—such as coherence or coreference—rather than providing a comprehensive and structured assessment of the full range of discourse phenomena. This narrow scope limits their ability to serve as stand-alone measures of document-level translation quality, underscoring the need for reproducible human protocols that can capture pragmatic, referential, and thematic dimensions in an integrated framework.

The development of test suites targeting specific discourse elements also represents an efficacious research methodology. This includes domain-specific investigations (Vojtěchová et al. 2019; Biçici 2019; Mukherjee and Yadav 2024; Bhattacharjee et al. 2024; Rozanov et al. 2024; Bawden and Sagot 2023), examinations involving linguistic features (Avramidis et al. 2019; Popović 2019; Raganato et al. 2019; Zouhar et al. 2020; Macketanz et al. 2021; Manakhimova et al. 2023; Savoldi et al. 2023; Ármannsson et al. 2024; Friðriksdóttir 2024; Manakhimova and Macketanz 2024; Dawkins et al. 2024), and analyses incorporating discourse features (Rysová et al. 2019; Kocmi et al. 2020; Avramidis et al. 2020; Scherrer et al. 2020; Mukherjee and Shrivastava 2023). Such studies underscore the capacity of discourse-sensitive evaluations to identify errors that sentence-level metrics tend to systematically overlook.

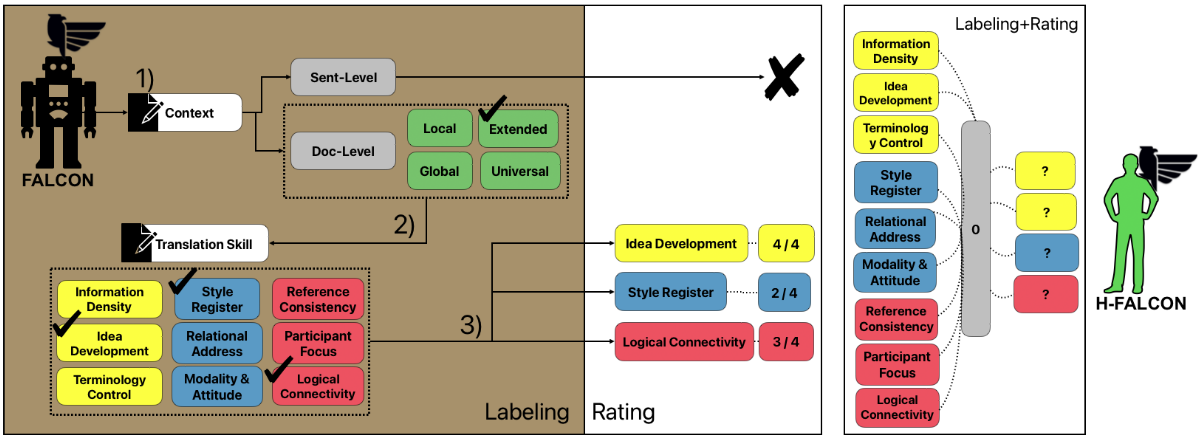


Figure 1: The evaluation process of FALCON consisting of labeling 1) relevant context knowledge and 2) assessment of translation skills, accompanied by 3) rating. This dual-phase process is integrated in H-FALCON by simultaneously conducting labeling and rating for all sentences.

2.3 The FALCON Framework

FALCON (Functional Assessment of Language and Contextuality in Narratives; Kim 2025) introduces a structured approach to document-level MT evaluation by integrating pragmatic and discourse-level considerations into a holistic scoring procedure. It is built on two central hypotheses:

(a) Document-level evaluation can be operationalized at the sentence level if contextual information is effectively transferred between sentences.

(b) This transferred information can be extracted exclusively from the source sentence, independent of the target language.

The framework organizes discourse phenomena into three meta-categories (Mode, Tenor, Field) and nine sub-categories (Information Density, Idea Development, Terminology Control, Style Register, Relational Address, Modality & Attitude, Reference Consistency, Participant Focus, and Logical Connectivity), collectively referred to as ”translation skills.” For each sentence, the judge selects the three most relevant skills that apply regardless of the target language. Constraining the selection to three choices is to ensure better stability of the model outputs. Before this task, sentences deemed not to require context are filtered out via ”labeling,” where annotators classify context knowledge into five predefined types: Sentence-level, Local, Extended, Global, or Universal. In the subsequent ”rating” stage, the judge assigns a 4-point score to each designated skill. This process is depicted in Figure 1. These scores can then be averaged per segment or per skill set to yield interpretable indicators of document-level performance.

Designed primarily for LLM-as-judge environment, FALCON handles document context by presenting two preceding and two following segments from both the source and reference translations. For its human evaluation, it adopts the standard document-level protocol described in Kocmi et al. (2024). Nonetheless, the dependence of FALCON on partial context prompts inquiries into the completeness with which the model apprehends holistic discourse phenomena.

These concerns are addressed through the testing of their structured categorization using human evaluation, alongside the meta-evaluation of FALCON informed by their judgments. The empirical insights derived from these studies have assisted in finalizing H-FALCON, an effective and reliable protocol for human evaluation.

3 Human Evaluation Setup

We conduct human evaluation employing the original framework of FALCON. This evaluation encompasses two primary tasks: i) labeling contextual knowledge, and ii) rating three pertinent translation skills.

Evaluation I: Context Knowledge Judges determine the context level required for accurate translation in five categories with domain information, assuming some sentences need no additional context.

Evaluation II: Translation Skill Judges choose

Domain	**#. Seg**	**#. Sent**	**Avg.**	**Min.**	**Max.**
Overall	809	16.64	1	1.15	61
Literary	76	15.63	2	1.34	41
News	233	24.27	3	1.16	60
Social	500	13.24	1	1.11	55

Table 1: Overall and per-domain statistics for the evaluation dataset. The size of segments (**#. Seg**) and the mean number of sentences (**#. Sent**) per segment are computed. Additionally, the word size is assessed in terms of average, minimum, and maximum values.

Group	**J$_1$-J$_2$**	**J$_2$-J$_3$↑**	**J$_3$-J$_1$**
Original	0.146	0.500	0.117
Local + Sent-level	0.116	**0.580**	0.103
Extended + Universal	0.145	0.568	0.124
Global + Sent-level	0.143	0.502	0.117
Global + Universal	0.146	0.501	0.117
Extended + Local	**0.158**	0.500	**0.143**
Extended + Global	0.148	0.499	0.117
Global + Local	0.149	0.499	0.117
Extended + Sent-level	0.152	0.480	0.097
Local + Universal	0.152	0.463	0.133
Sent-level + Universal	0.136	0.418	0.087

Table 2: Cohen's kappa scores (κ) representing pairwise agreements for the context knowledge labels after consolidation two categories. For anonymity, the evaluators are denoted as J_1, J_2, and J_3.

the top three skills from nine categories required for translation.

3.1 Dataset

We utilize the WMT24++ (Deutsch et al., 2025) English-to-Korean translation dataset for our evaluation. This dataset includes 1,000 segments across four domains: social, news, speech, and literary; with translations from 10 different systems for analysis. Ineffective models with poor translations are excluded to maintain context sensitivity. We exclude the speech domain as segments are inherently contextless. Our dataset aims to maintain similar domain proportions while respecting document boundaries. Given the greater length of literary texts, consecutive sentences are clipped to fit the desired length. Irrelevant elements like hyperlinks, hashtags, and timestamps are excluded, but emojis and user tags remain.

To create a genuine sentence-level environment, the original data is segmented into sentences that warrant evaluation and seamlessly align with the target side, using NLTK (Bird et al., 2009) for English and KSS (Korean Sentence Splitter)[1] for Korean. The alignment of the source and target segments is initially performed automatically by taking into account new line indicators, following which we manually verify their validity. A total of 809 unique sentences are used for evaluation, as detailed in Table 1.

3.2 Recruitment & Orientation

Three professional translators, all native speakers of Korean with 5 to 10 years of experience in translating English, were engaged in this study. For confidentiality, they were anonymized as J_1, J_2, and J_3, and collectively referred to as judges or evaluators. Given the shorter evaluation unit compared to the original dataset, we anticipated that a judge would handle on average 60 sentences an hour, amounting to 13.3 hours per task and totaling 27 hours for each judge across two tasks. The remuneration was $30 per hour.

Orientation was conducted online via Google Meet to explain evaluation guidelines and demonstrate the platform. Participants accessed the platform during the session to conduct a preliminary evaluation. They had one week to complete the evaluation. As time was tracked per item, they were advised to focus on the task and avoid distractions. Participants were granted unrestricted access to the entire document and were permitted to review and amend their evaluations at their discretion prior to final submission.

3.3 Platform & Interface

Label Studio[2] serves as the primary platform for our evaluation, with its interface depicted in Figure 6 in the appendix. Evaluators are consistently able to consult the definitions of labels along with prior annotations of segments and pertinent domain information.

3.4 Metrics

IAA is the main metric for assessing platform reliability. In Evaluation I, Cohen's Kappa κ is calculated for each judge using Equation 1, where P_o is the observed agreement between judges, and P_e is the chance agreement based on marginal label distributions.

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{1}$$

[1] `https://github.com/hyunwoongko/kss`

[2] `https://labelstud.io`

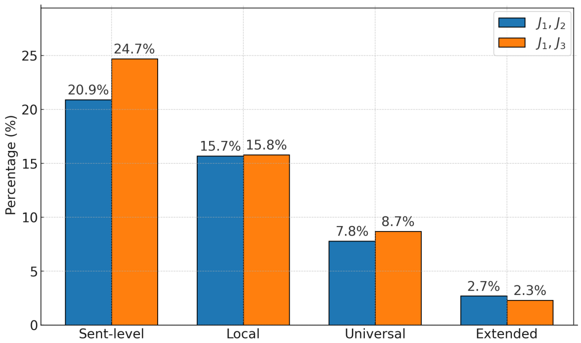


Figure 2: The label distribution shows **disagreement** of context knowledge involving J_1, mainly occurring at the SENTENCE-LEVEL between the two judges.

For a multi-label annotation in Evaluation II, Jaccard Similarity J is used as shown in Equation 2, with A and B being the label sets from two annotators. For qualitative assessment, participant feedback is gathered via Google Sheets, and subsequent linguistic analysis is performed on our part.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

4 Result

4.1 Evaluation I - fair to moderate agreement

Table 2 shows quite low pairwise agreement from 0.365 (J_3 and J_1) to 0.499 (J_2 and J_3), indicating a fair to moderate agreement level (Landis and Koch, 1977). The notably lower correlation observed when J_1 participates suggests that J_1 may adhere to a different evaluative standard compared to the other judges. To clarify this issue, we analyze J_1's disagreement rate with others. As in Figure 2, J_1 frequently disagrees on SENTENCE-LEVEL, differing from J_2 and J_3 by 20.9% and 24.7%, respectively, while both judges prefer the LOCAL for this case at 19.0% and 24.1%.

A qualitative study of their feedback reveals that **pronouns often cause this disagreement of SENTENCE-LEVEL and LOCAL context knowledge.** For example, when translating the pronoun ”it” into an equivalent pronoun in the target language, and this action is judged to be correct, the LOCAL context is selected. Conversely, if one deems this translation approach incorrect and suggests considering hypothetical instances where the actual noun is used in translation, they would label it LOCAL. An illustrative example is presented in Table 3. J_2 states that ”the interpretation of a pronoun's referent also influences verb choice, thus I categorize the sentence as LOCAL.”

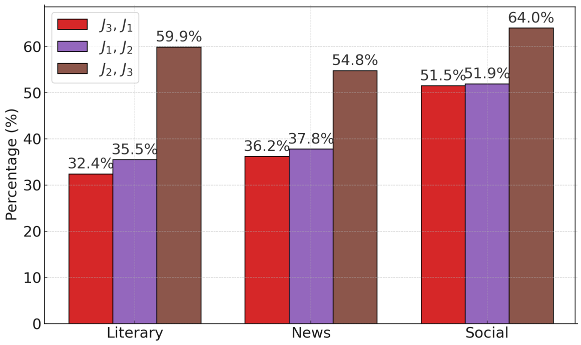


Figure 3: The distribution of labels indicating **agreement** regarding translation skills as evaluated by two judges across three domains. Judges agree most in the social domain.

4.2 Evaluation II - moderate to substantial agreement

The judges have demonstrated an average Jaccard Similarity of 50.8%, with each pair achieving consistent labeling agreement percentages of 46.3%, 61.0%, and 45.3%, respectively demonstrating moderate to substantial concordance as per empirical Jaccard-based standards (Zhang and Zhou 2014; Rajpurkar et al. 2016). Consistent with the prior analysis, the average agreement remains relatively low, with the most substantial concordance observed in the J_2 and J_3 pair.

Upon analysis by domain as presented in Figure 3, the greater participant agreement is observed in the social domain. Furthermore, the agreed-upon labels as calculated between J_2 and J_3 in Figure indicate that the most frequently concurred top three labels fall under the meta-category of TENOR, specifically STYLE REGISTER, RELATIONAL ADDRESS, and MODALITY & ATTITUDE. These findings suggest that

SRC	I bought it like that and couldn't modify it, so I had to design around it.
TGT	구매했을 때부터 그런 형태였고, 수정할 수 없어서 그 형태에 맞춰 디자인해야 했어요.
BT	*It was in that form from the moment I purchased it, and since I couldn' t change it, I had to design everything to fit that shape.*

Table 3: A notable instance provoking frequent misunderstanding between SENTENCE-LEVEL and LOCAL labels due to pronouns. The source (SRC) and target (TGT) segments are exemplified with the help of back-translation (BT).

Group	Pair	acc (%)↑	κ
H vs. H	J_2, J_3	**70.09**	**0.4995**
	J_1, J_2	66.25	0.3883
	J_1, J_3	62.92	0.3646
H vs. M	J_3, `o4-mini`	**53.89**	**0.2535**
	J_1, `o4-mini`	52.29	0.1788
	J_2, `o4-mini`	51.67	0.1891
	J_3, `o3`	51.17	0.2059
	J_1, `o3`	50.31	0.1484
	J_2, `o3`	49.57	0.1591
	J_1, `4.1-mini`	42.77	0.0802
	J_2, `4.1-mini`	39.80	0.0478
	J_3, `4.1-mini`	39.68	0.0750
M vs. M	`o3`, `o4-mini`	**71.69**	**0.5239**
	`4.1-mini`, `o4-mini`	47.22	0.2046
	`o3`, `4.1-mini`	40.30	0.1068

Table 4: Pairwise accuracy and Cohen's Kappa κ by human-human (H-H), human-model (H-M), and model-model (M-M) for context knowledge.

domains like social often require specific skills related to Tenor, leading to higher IAA.

In the feedback, J_1 articulates that there are certain skills that remain consistently significant across domains, citing Modality & Attitude within the literature and news domains, alongside Relational Address in the social domain. J_2 further asserts that each domain possesses inherent default skills, if not having any other urgent skills. While agreeing with this opinion, J_3 asserts that this makes such domains relatively straightforward to annotate. They add that, unlike such a case, the news domain necessitates a broader range of skills, rendering it challenging to select only three critical skills.

5 Meta-Evaluation of FALCON

The human IAA attained in the current configuration is approximately $\kappa = 0.50$ for Evaluation I and $J = 0.61$ for Evaluation II. This section examines the efficacy of FALCON in its role as LLM-as-judge, making use of a broader array of more recent proprietary models in OpenAI, specifically `gpt-o3`, `gpt-o4-mini` and the originally used `gpt-4.1-mini`. The performance of these models is assessed using the same metrics as in § 3.4. Concurrently, we employ three gold standards that we have obtained to calculate the average pairwise accuracy for Evaluation I and the Micro F1 score for Evaluation II.

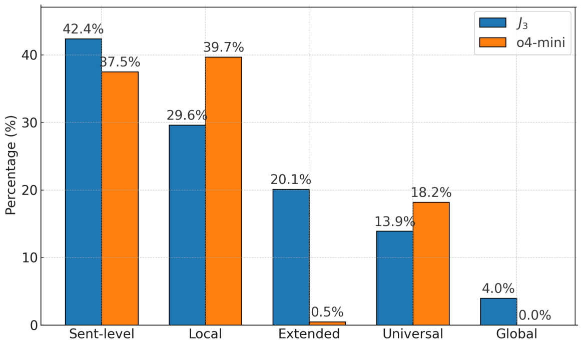


Figure 4: The label distribution shows **disagreement** of context knowledge between J_3 and `o4-mini`, the most agreed pair.

5.1 Meta-Evaluation I - fair at best with 53.8% of accuracy

Table 4 shows the pairwise accuracy and IAA in three pairs: human-human, human-model, and model-model. The model attained a maximum accuracy of 53.89% with `o4-mini`, and no model has reached an accuracy level comparable to that of the optimal human pair (J_2-J_3). Such performance remains inferior to the lowest performance among human pairs (J_1-J_3, 62.92%). Furthermore, the concordance with human evaluative standards is, at most, fair ($\kappa = 0.25$ by `o4-mini`), indicating a deficiency in the model's ability to effectively differentiate between labels in a manner comparable to human evaluators.

An analysis is subsequently conducted to ascertain the labels that most significantly contribute to the observed discrepancies. Figure 4 displays the distribution of labels upon which J_3 and `o4-mini`, the pair with the highest consensus, diverge. Notably, the largest portion of the disagreement is associated with Sentence-level, comprising 42.4% and 37.5%, respectively. Collectively, this analysis highlights a pronounced tendency for Sentence-level and Local labels to be primary drivers of disagreement. Furthermore, a domain-specific analysis reveals that 65.2% of these discrepancies occur within the social domain.

5.2 Meta-Evaluation II - fair with 53.6% of accuracy

The pairwise agreement scores illustrated in Table 5 indicate that the most optimal human-model agreement occurs with `o4-mini` ($J = 0.406$), which is considerably inferior to the agreement levels observed between human-human or model-model pairings. Furthermore, the precision of the model is

Group	Pair	avg. $J\uparrow$	f1
H vs. H	J_2–J_3	**0.6098**	**0.7183**
	J_1–J_2	0.4629	0.5915
	J_1–J_3	0.4529	0.5737
H vs. M	J_2, `o4-mini`	**0.4067**	**0.5360**
	J_3, `o4-mini`	0.3976	0.5272
	J_2, `o3`	0.3970	0.5231
	J_1, `o4-mini`	0.3912	0.5196
	J_1, `o3`	0.3829	0.5099
	J_2, `4.1-mini`	0.3704	0.4931
	J_3, `4.1-mini`	0.3683	0.4893
	J_1, `4.1-mini`	0.3660	0.4871
	J_3, `o3`	0.3625	0.4854
M vs. M	`o3`, `o4-mini`	**0.5972**	**0.7082**
	`4.1-mini`, `o4-mini`	0.4665	0.5948
	`4.1-mini`, `o3`	0.4250	0.5554

Table 5: Average pairwise Jaccard Similarity J and Micro F1 by human-human (H-H), human-model (H-M), and model-model (M-M) for translation skills.

quantified at 53.6%, akin to the earlier task. Notably, the agreement levels among models are comparable to those observed in human-human pairings, achieving similar levels at best $J = 0.597$. Such results imply that while models exhibit consistency in their predictions, they might rely on underlying patterns distinct from human reasoning processes. These observations underscore the necessity to enhance model accuracy and align their performance more closely with human evaluative criteria.

6 New Methodology: H-FALCON

6.1 Background

Our comprehensive experiments manifest that evaluating discourse phenomena in the current protocol of FALCON poses a significant challenge for both human evaluators and models. For human evaluators, the objectives of translation lack clear delineation. Certain competencies are consistently required, while others are not invariably necessary. Moreover, the determination of the requisite number of skills per sentence should exhibit flexibility, contingent upon the domain characteristics, target language, and the sentence length. This inherent instability has led to a diminished reliability of the LLM-as-judge, causing low reliability whereby one of the two answers may be incorrect. This situation underscores the critical necessity for enhancements to the existing evaluation framework. In pursuit of this objective, we propose a novel methodology H-FALCON grounded in two hypotheses:

Hyp 1. Every sentence needs context.

Hyp 2. The judges have the discretion to select the number of translation skills.

6.2 Design

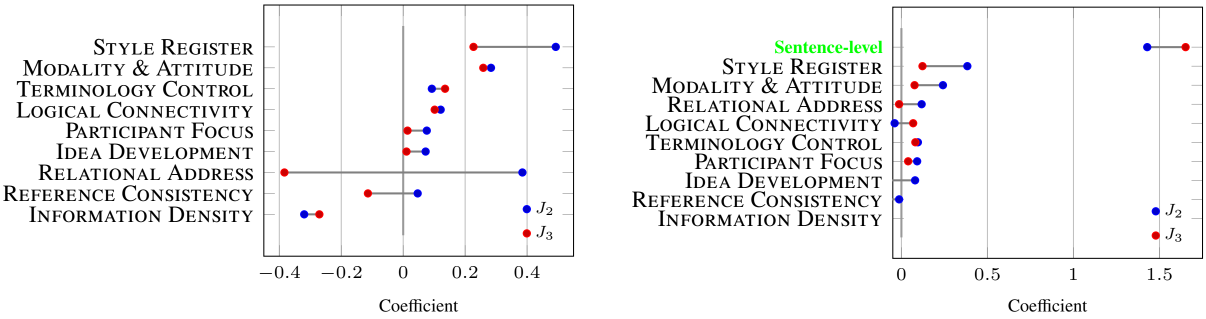
The fundamental concept of H-FALCON involves the integration of skill annotation with assessment while ensuring consideration of all sentences in the evaluation, as illustrated in Figure 1. Each sentence initially has all skills labeled as NOT RELEVANT. Then, judges are required to assess *only* the pertinent skills using a triadic scale—HIGH, MEDIUM, and LOW—consistent with Julian House's theoretical framework (House, 2015), thereby conducting the labeling and rating processes concurrently. The preliminary feedback from our human evaluators agrees with a reduced scale from four to three points, indicating that discourse phenomena frequently lend themselves to relatively clear judgments.

6.3 Experiment Setup

A selection of 300 new subsamples is extracted from WMT24++ (Deutsch et al., 2025), maintaining proportional representation across domains. Judge 2 and Judge 3 are hired again for this experiment. The evaluation environment remains consistent, with the same platform as in Figure 7 in the appendix; however, the task now specifically pertains to rating. The reliability of label selection is measured by excluding NOT RELEVANT labels and calculating Jaccard Similarity between the judges. The Pearson, Spearman, and Kendall's tau correlation coefficients are calculated from the scores.

In addition to this, traditional evaluations of MQM-style errors at the sentence level are undertaken using a 4-point scale, while a holistic score encompassing both sentence and document levels is solicited on a 10-point scale. Utilizing this data, our objective is to ascertain the benchmark for an acceptable IAA threshold applicable to H-FALCON and to elucidate the relationships among the three evaluation metrics. All ratings are conducted simultaneously on an individual sentence basis. The use of varying scales is intentionally employed to mitigate the risk of task confusion.

To investigate the extent to which sentence-level and document-level scores predict the holistic score, we fit multiple linear regression models. For each



(a) Doc-level Features

(b) Doc- & Sent-level Features

Figure 5: Linear regression coefficients for J_2 and J_3 with (right) and without (left) sentence-level score. Features with scores near 0 have minimal influence on the holistic score.

Type	Pearson	Spearman	Kendall
Sentence-level	0.494	0.441	0.413
H-FALCON (sum)	0.499	0.483	0.378
H-FALCON (count)	0.562	0.545	0.486
Holistic	0.650	0.587	0.502

Table 6: Correlations between two raters for sentence-level, document-level (of two different styles), and holistic scores.

	Doc		**Doc + Sent**	
Metric	J_2	J_3	J_2	J_3
R^2	0.1195	0.0865	0.4700↑	0.6093↑
Intercept	6.4593	7.7551	2.0968	2.4793

Table 7: The explanatory power (R^2) of models when with document-level score (**Doc**) and with document- and sentence-level scores (**Doc+Sent**), with intercept.

rater, we use the holistic score as the dependent variable and the individual component scores as independent variables, both with an intercept term. This approach allows us to quantify the contribution of each component while controlling for the influence of the others.

6.4 Analysis of Reliability

The calculated Jaccard Similarity index between the two judges is 0.532, which remains consistently low, similar to the outcomes observed in the preceding experiment. This result affirms the inherent challenge in attaining consensus, irrespective of the methodology employed for label collection. Subsequently, each selected label is attributed a numerical value, specifically (HIGH=3, MEDIUM=2, LOW=1, NOT RELEVANT=0), in order to generate a document-level score. The values are either aggregated (”sum”) or quantified (”count”). Upon comparing the correlations of sentence, document, and holistic scores between the judges, Table 6 demonstrates that **the document-level scoring scheme has achieved a level of agreement comparable to that of sentence-level scores, thereby demonstrating its validity**. Calculating the number of errors has resulted in a somewhat improved agreement, highlighting the potential as an effective method for annotating document-level scores.

6.5 Analysis of Scoring System

Upon the validity of the document-level score, Figure 5 displays regression coefficients indicating the direction and magnitude of each component's effect on the holistic score. In Figure 5-a), the positive coefficients for STYLE REGISTER (0.46), RELATIONAL ADDRESS (0.38), and MODALITY AND ATTITUDE (0.24) suggest that **aspects of tone, interpersonal engagement, and stylistic choices are more influential on holistic score than purely content-oriented measures**. The negative impact of INFORMATION DENSITY may indicate excessive content compression, compromising overall quality.

Notwithstanding the influence observed, the low R^2 in Table 7 indicates that document-level scores alone are insufficient in fully determining the holistic score. Incorporation of the sentence-level score enhances the model's explanatory power to a satisfactory degree, as evidenced by $R^2 = 0.6$ and corroborated in Figure 5-b) with a coefficient of 1.65 (J_3). This further substantiates that **the sentence-level score predominantly drives the variance in the holistic score.** The more pronounced R^2 for J_3 implies a more uniform weighting of these features in comparison to J_2.

Moreover, each sentence is annotated with at least one translation skill by both evaluators. Based on these findings, we reject the claim that only a subset of sentences requires contextual information (Castilho, 2022). In contrast, we assert that contextual information is crucial for every sentence, although its impact on the total score is relatively minimal. In particular, our experiment shows that approximately 10.5% of the discourse phenomena contribute to the overall score. Nonetheless, this does not diminish the importance of document-level evaluation, as this will serve as the distinguishing factor for models demonstrating superior performance.

7 Conclusion

Our findings challenge prevailing assumptions in MT evaluation by showing that contextual information, while modest in magnitude, is universally present and measurably influences human judgments. By operationalizing this insight through H-FALCON, we offer a reproducible, context-aware evaluation protocol that attains stronger alignment with human preferences than traditional sentence-level approaches. This work underscores the necessity of shifting from narrow, sentence-bounded metrics toward richer, document-level assessments that better capture the pragmatic realities of translation quality. As MT performance converges at the sentence level, such holistic, context-sensitive evaluation will be critical for driving the next phase of progress in the field.

8 Limitation

The scope of this research is confined to a single mid-resourced language pair. This is deemed acceptable as our investigation is concentrated on the human evaluation environment, which remains consistent across various languages. Nevertheless, the performance and reliability of FALCON may exhibit variability per language pair. In the human evaluation, only three annotators are engaged, one of whom demonstrates notably divergent behavior. Similarly, the consensus on skill selection in § 6.4 remains unenhanced even with new methodology. These observations suggest that a more proactive pre-meeting should have been organized.

References

David Anugraha, Garry Kuwanto, Lucky Susanto, Derry Tanti Wijaya, and Genta Winata. 2024. MetaMetrics-MT: Tuning meta-metrics for machine translation via human preference calibration. In *Proceedings of the Ninth Conference on Machine Translation*, pages 459–469, Miami, Florida, USA. Association for Computational Linguistics.

Eleftherios Avramidis, Vivien Macketanz, Aljoscha Burchardt, and et al. 2020. Fine-grained linguistic evaluation for state-of-the-art machine translation. *arXiv preprint arXiv:2010.06359*.

Eleftherios Avramidis, Vivien Macketanz, Ulrich Strohriegel, and Aljoscha Burchardt. 2019. Linguistic evaluation of german-english machine translation using a test suite. *arXiv preprint arXiv:1910.07457*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Rachel Bawden and Benoît Sagot. 2023. RoCS-MT: Robustness challenge set for machine translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 198–216, Singapore. Association for Computational Linguistics.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Soham Bhattacharjee, Biswajit Gain, and Asif Ekbal. 2024. Domain dynamics: Evaluating large language models in english-hindi translation. In *Proceedings of the Ninth Conference on Machine Translation (WMT24)*. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Ergun Biçici. 2019. Machine translation with parfda, moses, kenlm, nplm, and pro. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 66–73. Association for Computational Linguistics.

Sheila Castilho. 2022. How much context span is enough? examining context-related issues for document-level MT. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3017–3025, Marseille, France. European Language Resources Association.

Nicolas Dahan, Rachel Bawden, and François Yvon. 2024. Survey of automatic metrics for evaluating machine translation at the document level. Technical report, HAL Open Science. Available at HAL Open Science.

Hillary Dawkins, Isar Nejadgholi, and Chi-Kiu Lo. 2024. WMT24 test suite: Gender resolution in speaker-listener dialogue roles. In *Proceedings of the Ninth Conference on Machine Translation*, pages 307–326, Miami, Florida, USA. Association for Computational Linguistics.

Daniel Deutsch, Eleni Briakou, Isaac Caswell, Max Finkelstein, Roni Galor, and 1 others. 2025. WMT24++: Expanding the language coverage of wmt24 to 55 languages & dialects. *arXiv preprint arXiv:2502.12404*.

Sigríður Rut Friðriksdóttir. 2024. The genderqueer test suite. In *Proceedings of the Ninth Conference on Machine Translation (WMT24)*, pages 265–273. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Christian Hardmeier, Liane Guillou, Pierre Lison, and Jörg Tiedemann. 2015. Report on the discomt 2015 shared task on discourse translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16. ACL.

Juliane House. 2015. *Translation Quality Assessment: Past and Present*. Routledge, London and New York.

Zheng Jiang, Yang Yu, Yang Feng, Bing Qin, and Ting Liu. 2022. Blonde: An automatic evaluation metric for document-level natural language generation. In *Proceedings of NAACL*, pages 1679–1698.

Prathyusha Jwalapuram, Barbara Rychalska, Shafiq Joty, and Dominika Basaj. 2021. Dip benchmark tests: Evaluation benchmarks for discourse phenomena in {mt}.

Ahrii Kim. 2025. Falcon: Holistic framework for document-level machine translation evaluation. *TechRxiv*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and 3 others. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, Maja Popović, and Mariya Shmatova. 2022. Findings of the 2022 conference on machine translation (wmt22). In *Proceedings of the Seventh Conference on Machine Translation*, pages 1–45, Abu Dhabi. Association for Computational Linguistics.

Tom Kocmi, Tomasz Limisiewicz, and Gabriel Stanovsky. 2020. Gender coreference and bias evaluation at wmt 2020. *arXiv preprint arXiv:2010.06018*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional quality metrics (mqm): A framework for defining translation quality. In *Proceedings of the Eighth Conference on International Language Resources and Evaluation (LREC'14)*, pages 1285–1291. European Language Resources Association (ELRA).

Vivien Macketanz, Eleftherios Avramidis, and Aljoscha Burchardt. 2021. Linguistic evaluation for the 2021 state-of-the-art machine translation systems for german to english and english to german. In *Proceedings of the Sixth Conference on Machine Translation (WMT21)*, pages 1122–1137. Association for Computational Linguistics.

Sabina Manakhimova, Eleftherios Avramidis, and Vivien Macketanz. 2023. Linguistically motivated evaluation of the 2023 state-of-the-art machine translation: Can chatgpt outperform nmt? In *Proceedings of the Eighth Conference on Machine Translation (WMT23)*. Association for Computational Linguistics.

Sabina Manakhimova and Vivien Macketanz. 2024. Investigating the linguistic performance of large language models in machine translation. In *Proceedings of the Ninth Conference on Machine Translation (WMT24)*. Association for Computational Linguistics.

Anwesha Mukherjee and Manish Shrivastava. 2023. Iiit hyd's submission for wmt23 test-suite task. In *Proceedings of the Eighth Conference on Machine Translation (WMT23)*. Association for Computational Linguistics.

Anwesha Mukherjee and Shruti Yadav. 2024. Cost of breaking the llms. In *Proceedings of the Ninth Conference on Machine Translation (WMT24)*. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. ACL.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2019. Evaluating conjunction disambiguation on english-to-german and french-to-german wmt 2019 translation hypotheses. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 597–602. Association for Computational Linguistics.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2019. The mucow test suite at wmt 2019: Automatically harvested multilingual contrastive word sense disambiguation test sets for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 603–611. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2383–2392. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nikita Rozanov, Vladislav Pankov, and Danila Mukhutdinov. 2024. Isochronometer: A simple and effective isochronic translation evaluation metric. *arXiv preprint arXiv:2410.11127*.

Kateřina Rysová, Magdaléna Rysová, Tomáš Musil, Lucie Poláková, and Ondřej Bojar. 2019. A test suite and manual evaluation of document-level NMT at WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 455–463, Florence, Italy. Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2023. Test suites task: Evaluation of gender fairness in mt with must-she and ines. *arXiv preprint arXiv:2310.19345*.

Yves Scherrer, Alessandro Raganato, and Jörg Tiedemann. 2020. The mucow word sense disambiguation test suite at wmt 2020. In *Proceedings of the Fifth Conference on Machine Translation (WMT20)*.

Himanshu Sharma and Bharat Ram Sridhar. 2025. Document-level machine translation through discourse modelling: A survey. *CFILT - IITB*.

Katherine Thai, Magdalena Karpinska, Kalpesh Krishna, Baishakhi Ray, Kathleen McKeown, Ron Artstein, and Benjamin Van Durme. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1256–1274, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Paola Valli. 2015. The TAUS quality dashboard. In *Proceedings of the 37th Conference Translating and the Computer*, pages 127—136, London, UK. AsLing.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Tereza Vojtěchová, Matúš Novák, Matěj Klouček, and Ondřej Bojar. 2019. Sao wmt19 test suite: Machine translation of audit reports. *arXiv preprint arXiv:1909.01701*.

Min-Ling Zhang and Zhi-Hua Zhou. 2014. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26 (8):1819–1837.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

Wei Zhao, Michael Strube, and Steffen Eger. 2023. DiscoScore: Evaluating text generation with BERT and discourse coherence. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3865–3883, Dubrovnik, Croatia. Association for Computational Linguistics.

Vilém Zouhar, Tereza Vojtěchová, and Ondřej Bojar. 2020. Wmt20 document-level markable error exploration. In *Proceedings of the Fifth Conference on Machine Translation (WMT20)*, pages 347–356. Association for Computational Linguistics.

Björn Ármannsson, Hrafn Hafsteinsson, and Atli Jasonarson. 2024. Killing two flies with one stone: An attempt to break llms using english→icelandic idioms and proper names. *arXiv preprint arXiv:2410.03394*.

A Analysis of Collected Data

We report the number of annotations per context and translation skills of our human evaluation, as presented in Table 8-a). The data indicate that the distribution is consistent across the judges. The majority of judges have determined that approximately 60% of the data is translatable without context. However, there is significant variation in the specific sentences chosen for this label. Among the types of context-dependent knowledge, LOCAL constitutes the largest proportion, averaging 24%. It is also observed that there are almost no annotations for GLOBAL, suggesting that this type of context may not be quantifiable on a sentence-by-sentence basis.

With reference to translation skills as outlined in Table 8-b), STYLE REGISTER and RELATIONAL ADDRESS constitute some of the most frequently required skills, accounting for 21.23% and 18.47%, respectively. This finding, along with feedback from the judges, suggests that they are crucial skills for context-aware translation. Conversely, INFORMATION DENSITY emerges as the least selected skill, with a prevalence of approximately 2.6%. This may be attributed to either a deficiency in the judges' comprehension or the minimal significance attributed to the discourse. All in all, further elucidation is necessary.

B Descriptions of Context Levels

Sentence-level The sentence can be fully understood and translated without any outside information. All necessary meaning is present within the sentence itself — vocabulary, grammar, and semantics are straightforward.

Local Understanding requires minimal surrounding context — maybe the previous or next sentence — but nothing broader. Without it, pronouns, references, or logical connectors might be confusing.

Extended Grasping the meaning requires understanding the broader scene, paragraph, or emotional flow. Cultural nuance, emotional undertones, or evolving character perspectives start to matter.

Global The sentence depends on knowledge of the entire work (novel, article, movie) or even multiple entries (book series, TV seasons). Important world-building, character arcs, fictional

(a) Context Knowledge

Label	J_1	J_2	J_3	Avg↑
SENT-LEVEL	63.16	60.57	54.14	59.29
LOCAL	23.11	21.76	26.33	23.73
UNIVERSAL	10.88	13.23	10.01	11.37
EXTENDED	2.84	4.08	9.52	5.48
GLOBAL	0.00	0.37	0.00	0.12

(b) Translation Skills

Label	J_1	J_2	J_3	Avg↑
STYLE REGISTER	21.26	20.77	21.67	21.23
RELATIONAL ADDRESS	19.70	19.28	16.44	18.47
REFERENCE CONSISTENCY	13.51	14.50	17.84	15.28
MODALITY AND ATTITUDE	17.39	14.05	12.48	14.64
TERMINOLOGY CONTROL	10.47	7.95	8.24	8.89
IDEA DEVELOPMENT	5.07	7.62	7.70	6.80
PARTICIPANT FOCUS	4.04	8.82	6.55	6.47
LOGICAL CONNECTIVITY	5.40	4.49	6.84	5.58
INFORMATION DENSITY	3.17	2.51	2.22	2.63

Table 8: Proportion of context and translation skill labels annotated by three judges (%)

history, or long-term motifs influence meaning.

Universal Understanding draws on extensive external knowledge — history, philosophy, science, mythology, social structures, or famous world events. Without that shared knowledge, translation risks misfiring badly.

C Descriptions of Translation Skills

Information Density Does the sentence compress information into abstract or complex structures required by the genre or audience? Important linguistic devices are nominalization, complex noun phrases, embedded clauses, compounding, metaphors, analogies, symbolic imagery, etc.

Idea Development Do some elements in the sentence influence the development of the central theme and the rhetorical structure expected by the genre? Important linguistic devices are discourse markers, schematic structures (e.g., introduction-body-conclusion), paragraph transitions, etc.

Terminology Control Does the sentence have technical or domain-specific vocabulary that requires accurate and consistent use across an entire text? Important linguistic devices are technical nouns, specialized terminology, standard collocations, fixed expressions, etc.

Style Register Do some elements in the sentence require a degree of linguistic politeness and stylistic appropriateness suited to the context and purpose of the text? Important linguistic devices are lexical choice, pronoun usage, verb conjugation, discourse markers, euphemisms, idiomatic expressions, etc.

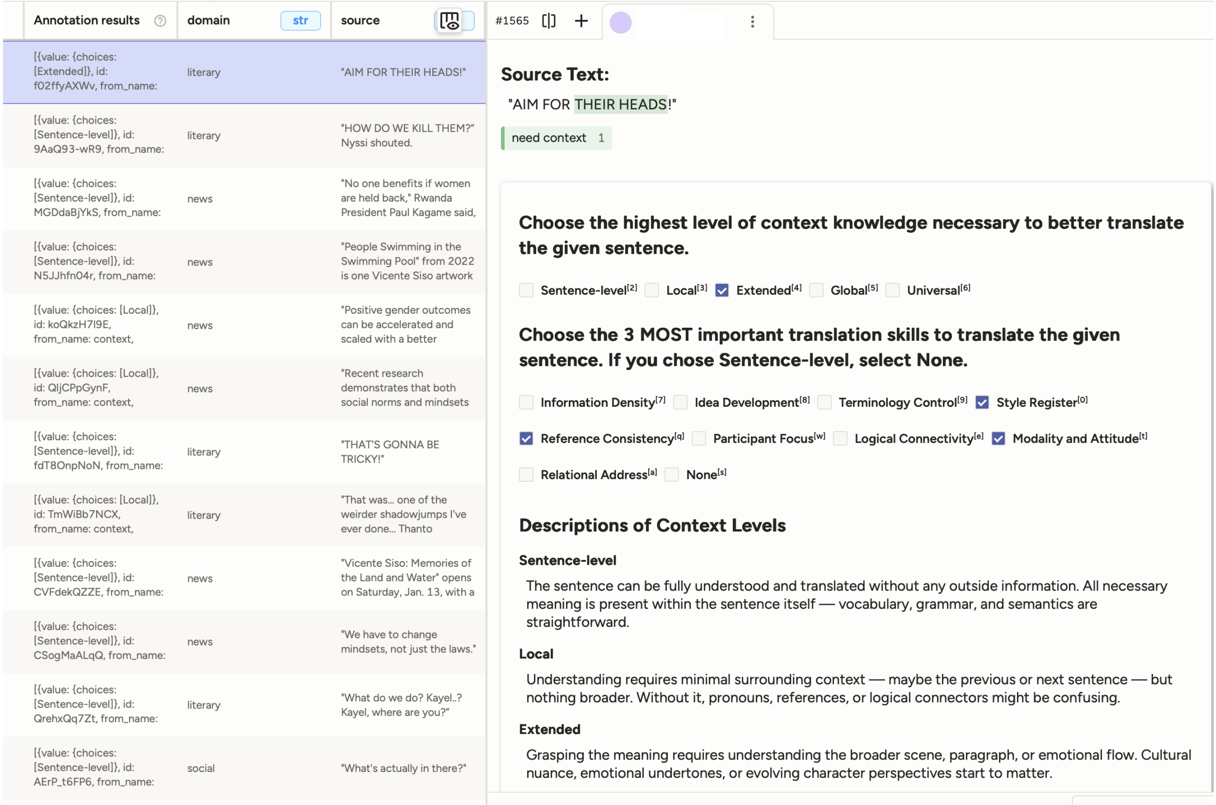
Reference Consistency Does the sentence contain elements that refer to the same entity within the text? The consistent use of such elements creates connections and coherence and ensures clear identification of participants, objects, and ideas throughout the text. Important linguistic devices are reference, substitution of clause, gender/tense/number agreement, deixis, ellipsis, repetition, synonyms, etc.

Logical Connectivity Does the sentence have connectors or structures that require clear expression of relationships — such as cause, contrast, or sequence — between ideas? Important linguistic devices are logical connectors (e.g., however, therefore), adversatives, causal linkers, etc.

Modality and Attitude Do some elements in the sentence express possibility, obligation, certainty, or speaker/writer's stance that convey the text's mood and tone? Important linguistic devices are modal verbs and auxiliaries (e.g., must, might), evaluative adjectives (e.g., important, unfortunate), stance adverbs (e.g., perhaps, clearly, surprisingly), emotionally charged expressions, subjunctive or conditional constructions, etc.

Relational Address Does the sentence rely on an understanding of the author's cultural, historical, or social background that affects his/her voice, intent, and the nuanced relationships with listener/reader? Important linguistic devices are gendered forms, titles and vocatives, pronoun, honorifics, relational expressions, sociolect, etc.

Participant Focus Should the emphasis of the sentence on key participants or elements (such as people, places, or objects) be preserved to convey the original meaning across a text? Important linguistic devices are subject-specific terminology, transitivity structures (verb types, selection of active/passive, selection of grammatical subject, use of nominalization instead of verb), etc.



Annotation results	domain	source
[{value: {choices: [Extended]}, id: f02ffyAXWv, from_name:	literary	"AIM FOR THEIR HEADS!"
[{value: {choices: [Sentence-level]}, id: 9AaQ93-wR9, from_name:	literary	"HOW DO WE KILL THEM?" Nyssi shouted.
[{value: {choices: [Sentence-level]}, id: MGDdaBjYkS, from_name:	news	"No one benefits if women are held back," Rwanda President Paul Kagame said,
[{value: {choices: [Sentence-level]}, id: N5JJhfn04r, from_name:	news	"People Swimming in the Swimming Pool" from 2022 is one Vicente Siso artwork
[{value: {choices: [Local]}, id: koQkzH7I9E, from_name: context,	news	"Positive gender outcomes can be accelerated and scaled with a better
[{value: {choices: [Local]}, id: QljCPpGynF, from_name: context,	news	"Recent research demonstrates that both social norms and mindsets
[{value: {choices: [Sentence-level]}, id: fdT8OnpNoN, from_name:	literary	"THAT'S GONNA BE TRICKY!"
[{value: {choices: [Local]}, id: TmWiBb7NCX, from_name: context,	literary	"That was... one of the weirder shadowjumps I've ever done... Thanto
[{value: {choices: [Sentence-level]}, id: CVFdekQZZE, from_name:	news	"Vicente Siso: Memories of the Land and Water" opens on Saturday, Jan. 13, with a
[{value: {choices: [Sentence-level]}, id: CSogMaALqQ, from_name:	news	"We have to change mindsets, not just the laws."
[{value: {choices: [Sentence-level]}, id: QrehxQq7Zt, from_name:	literary	"What do we do? Kayel..? Kayel, where are you?"
[{value: {choices: [Sentence-level]}, id: AErP_t6FP6, from_name:	social	"What's actually in there?"

Source Text:

"AIM FOR THEIR HEADS!"

Choose the highest level of context knowledge necessary to better translate the given sentence.

□ Sentence-level[2] □ Local[3] ☑ Extended[4] □ Global[5] □ Universal[6]

Choose the 3 MOST important translation skills to translate the given sentence. If you chose Sentence-level, select None.

□ Information Density[7] □ Idea Development[8] □ Terminology Control[9] ☑ Style Register[0]

☑ Reference Consistency[q] □ Participant Focus[w] □ Logical Connectivity[e] ☑ Modality and Attitude[t]

□ Relational Address[a] □ None[s]

Descriptions of Context Levels

Sentence-level

The sentence can be fully understood and translated without any outside information. All necessary meaning is present within the sentence itself — vocabulary, grammar, and semantics are straightforward.

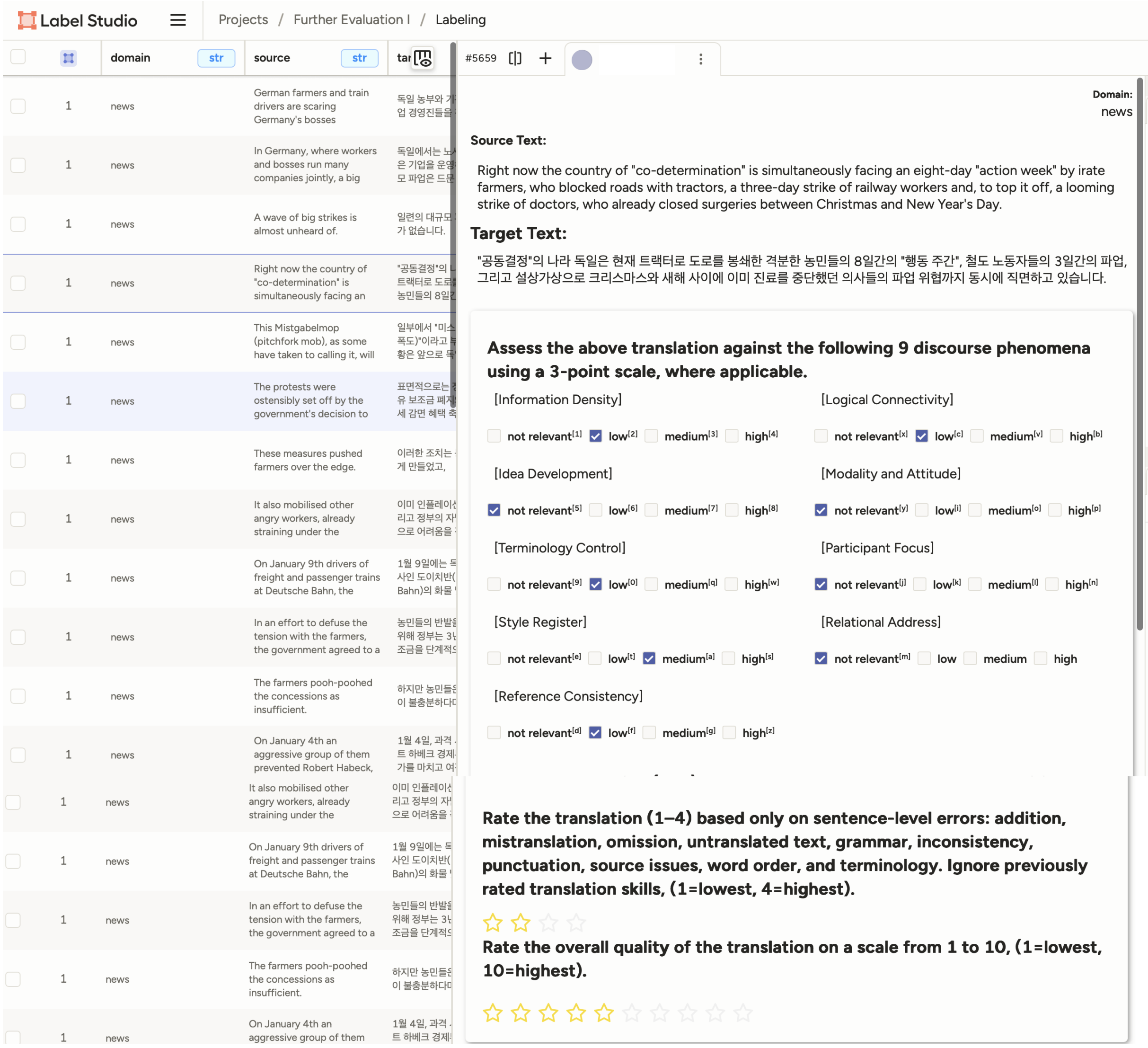
Local

Understanding requires minimal surrounding context — maybe the previous or next sentence — but nothing broader. Without it, pronouns, references, or logical connectors might be confusing.

Extended

Grasping the meaning requires understanding the broader scene, paragraph, or emotional flow. Cultural nuance, emotional undertones, or evolving character perspectives start to matter.

Figure 6: The interface of Label Studio for the human evaluation (FALCON]) of context knowledge and translation skills. Comprehensive explanations for each category are consistently presented on the expanded view of the screen.



☐		domain str	source str	tar
☐	1	news	German farmers and train drivers are scaring Germany's bosses	독일 농부와 업 경영진들을
☐	1	news	In Germany, where workers and bosses run many companies jointly, a big	독일에서는 은 기업을 운 모 파업은 드
☐	1	news	A wave of big strikes is almost unheard of.	일련의 대규모 가 없습니다.
☐	1	news	Right now the country of "co-determination" is simultaneously facing an	"공동결정"의 트랙터로 도로 농민들의 8일
☐	1	news	This Mistgabelmop (pitchfork mob), as some have taken to calling it, will	일부에서 "미스 폭도)"이라고 황은 앞으로
☐	1	news	The protests were ostensibly set off by the government's decision to	표면적으로는 유 보조금 폐지 세 감면 혜택
☐	1	news	These measures pushed farmers over the edge.	이러한 조치는 게 만들었고,
☐	1	news	It also mobilised other angry workers, already straining under the	이미 인플레이 리고 정부의 자 으로 어려움을
☐	1	news	On January 9th drivers of freight and passenger trains at Deutsche Bahn, the	1월 9일에는 사인 도이치반(Bahn)의 화물
☐	1	news	In an effort to defuse the tension with the farmers, the government agreed to a	농민들의 반발 위해 정부는 3 조금을 단계적
☐	1	news	The farmers pooh-poohed the concessions as insufficient.	하지만 농민들 이 불충분하다
☐	1	news	On January 4th an aggressive group of them prevented Robert Habeck,	1월 4일, 과격 트 하베크 경제 가를 마치고 여
☐	1	news	It also mobilised other angry workers, already straining under the	이미 인플레이 리고 정부의 자 으로 어려움을
☐	1	news	On January 9th drivers of freight and passenger trains at Deutsche Bahn, the	1월 9일에는 사인 도이치반(Bahn)의 화물
☐	1	news	In an effort to defuse the tension with the farmers, the government agreed to a	농민들의 반발 위해 정부는 3 조금을 단계적
☐	1	news	The farmers pooh-poohed the concessions as insufficient.	하지만 농민들 이 불충분하다
☐	1	news	On January 4th an aggressive group of them	1월 4일, 과격 트 하베크 경제

Domain:
news

Source Text:

Right now the country of "co-determination" is simultaneously facing an eight-day "action week" by irate farmers, who blocked roads with tractors, a three-day strike of railway workers and, to top it off, a looming strike of doctors, who already closed surgeries between Christmas and New Year's Day.

Target Text:

"공동결정"의 나라 독일은 현재 트랙터로 도로를 봉쇄한 격분한 농민들의 8일간의 "행동 주간", 철도 노동자들의 3일간의 파업, 그리고 설상가상으로 크리스마스와 새해 사이에 이미 진료를 중단했던 의사들의 파업 위협까지 동시에 직면하고 있습니다.

Assess the above translation against the following 9 discourse phenomena using a 3-point scale, where applicable.

[Information Density]

☐ not relevant[1] ☑ low[2] ☐ medium[3] ☐ high[4]

[Logical Connectivity]

☐ not relevant[x] ☑ low[c] ☐ medium[v] ☐ high[b]

[Idea Development]

☑ not relevant[5] ☐ low[6] ☐ medium[7] ☐ high[8]

[Modality and Attitude]

☑ not relevant[y] ☐ low[i] ☐ medium[o] ☐ high[p]

[Terminology Control]

☐ not relevant[9] ☑ low[0] ☐ medium[q] ☐ high[w]

[Participant Focus]

☑ not relevant[j] ☐ low[k] ☐ medium[l] ☐ high[n]

[Style Register]

☐ not relevant[e] ☐ low[t] ☑ medium[a] ☐ high[s]

[Relational Address]

☑ not relevant[m] ☐ low ☐ medium ☐ high

[Reference Consistency]

☐ not relevant[d] ☑ low[f] ☐ medium[g] ☐ high[z]

Rate the translation (1–4) based only on sentence-level errors: addition, mistranslation, omission, untranslated text, grammar, inconsistency, punctuation, source issues, word order, and terminology. Ignore previously rated translation skills, (1=lowest, 4=highest).

★★☆☆

Rate the overall quality of the translation on a scale from 1 to 10, (1=lowest, 10=highest).

★★★★★☆☆☆☆☆

Figure 7: The interface of Label Studio for the human evaluation of H-FALCON. By default, all translation skills are initially designated as ”not relevant.” Both sentence-level and holistic scores are gathered concurrently.